BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e111301
doi: 10.3897/biss.7.111301

OPEN ACCESS

Conference Abstract

Big Data for Beginners

Pieter Huybrechts ‡

‡ Research Institute for Nature and Forest, Brussels, Belgium

Corresponding author: Pieter Huybrechts (pieterhuy@gmail.com)

Received: 17 Aug 2023 | Published: 18 Aug 2023

Citation: Huybrechts P (2023) Big Data for Beginners. Biodiversity Information Science and Standards 7: e111301. https://doi.org/10.3897/biss.7.111301

Abstract

With the increasing amount of datasets being published and made available through global aggregators, such as the Global Biodiversity Information Facility (GBIF), new opportunities have opened to answer research questions that previously could not be considered. Techniques for large scale data integration offer benefits for the biodiversity research community (Heberling et al. 2021, Kays et al. 2020), profiting from the great and continuing efforts in data mobilisation and standardisation (such as Darwin Core, Wieczorek et al. 2012). These benefits include integrating several large data sources or enriching existing occurrence data with other information. Several commonly encountered barriers to large-scale use of biodiversity occurrence data exist. These include the lack of facilities for local storage of large and rapidly changing datasets, the computational power required for processing, unfamiliarity with existing toolsets, and insufficient resources to maintain big data infrastructure. These challenges are well documented in the context of high-throughput genomics (Marx 2013), and more recently in occurrence-based biodiversity research (for example Thessen et al. 2018).

However, while these hurdles and bottlenecks are very real, several of them have low cost of entry solutions. The aim of this presentation is to encourage the community to explore ambitious queries, to combine and examine all available data in its totality and to break down specific technical barriers, by providing a practical overview for researchers to maximise the power of large-scale data processing in their work.

While big data processing may seem daunting, tools accessible to users without a background in big data are available for both local workstations and cloud computing services that allow for scalable data processing at low cost, for instance Databricks

Community Edition or Apache Arrow. Using these resources, researchers can incorporate larger datasets into existing protocols, and by doing so, uncover patterns and insights that would be otherwise impossible to acquire using smaller subsets of the ever-expanding complex set that biodiversity occurrence data presents.

Keywords

data integration, biodiversity data

Presenting author

Pieter Huybrechts

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Heberling JM, Miller J, Noesgaard D, Weingart S, Schigel D (2021) Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118 (6). https://doi.org/10.1073/pnas.2018093118
- Kays R, McShea W, Wikelski M (2020) Born-digital biodiversity data: Millions and billions. Diversity and Distributions 26 (5): 644-648. https://doi.org/10.1111/ddi.12993
- Marx V (2013) The big challenges of big data. Nature 498 (7453): 255-260. https://doi.org/10.1038/498255a
- Thessen A, Poelen J, Collins M, Hammock J (2018) 20 GB in 10 minutes: a case for linking major biodiversity databases using an open socio-technical infrastructure and a pragmatic, cross-institutional collaboration. PeerJ Computer Science 4 https://doi.org/10.7717/peerj-cs.164
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS One 7 (1). https://doi.org/10.1371/journal.pone.0029715